Approved For Release 2000/08/25 : CIA-RDP57-00042A000200150015-6

APPENDIX-X

Encoding Relations Between Index Entries

Summary

In many searches, no ambiguity or uncertainity can arise when the scope of search is defined by a combination of entities, processes and attributes corresponding to various index entries. This is the case, when the very nature of the entities, processes, etc. permits little or no uncertainty as to their interrelationship. Oranges are imported to Boston from California and never to California from Boston. In general, however, a considerable degree of uncertainty is possible or even probable.

This appendix is concerned with various ways in which relationships stated in a document can be encoded for machine searching.

One possibility is to attach significance to the order of citation of the index entries but this involves undue complications with machines available at present or likely to be constructed in the near future. The most practical approach is to establish a system of role indicators. In practice these would be symbols that would be attached to index entries for the purpose of resolving uncertainty as to the relationships existing between them. In order to keep the coding system as simple as possible, it would probably be best to employ a small number of role indicators each of which has a broad general significance open to argument position in formal logic. For certain important roles, e.g., "raw material", "product","conditioning agency", it may be worthwhile to set up special role indicators having specific significance.

Introduction

As we have seen, the basic step in indexing any given document is to decide which objects, persons, processes, attributes, locations, etc., referred to by the document, are of interest in selecting and correlating the information contained in the document. As a consequence of this, policy decisions as to how indexing can be accomplished most advantageously, must take into account the purpose or purposes that a file of documents must serve. Previous discussion has also pointed out, that the effectiveness of presently available automatic equipment is greatly increased by appropriately encoding index entries. In particular, it is highly advantageous to employ a coding system which is so constructed that both specific and generic terminology can be used to define and conduct searching and correlating operations.

Construction of an effective code requires that a large mass of terminology be appropriately processed. Appendix IX presents procedures and techniques that we have developed so as to expedite the analyzing and encoding of scientific and technical terms.

Once the appropriate index entries have been set up and encoded, their punching in cards can then provide a file searchable by machine. As already noted, it is an essential feature of the new indexing system that all the entries pertaining to any one document- or, more precisely, to any one unit of information - are punched one after another in a single card, or a sequence of cards which acts as a unit as far as machine searching and selecting operations are concerned. This makes it possible to direct a search to any one index entry or to various combinations of the same.

Juxtaposition of index entries in a single card- or sequence of cards acting as a unit - indicates in our system that the various entries pertain to some one document (or unit of information). This simple relationship of belonging together may suffice to avoid ambiguity and uncertainty in defining and conducting many searches. Thus if we direct a search to the combination of entries "fire" "gasoline" "extinguish" "foam" we will probably select documents pertaining to the use of a foam to extinguish a gasoline fire, but conceivably we might also locate items concerned with extinguishing a fire involving gasoline in foam form. Such items would probably be few in number as gasoline in foam form is rarely produced, even experimentally.

All possible ambiguity can be eliminated in simple cases by setting up/appropriate convention. Thus, for example, we might establish some the following simple convention to indicate that a given compound has certain physical properties. The convention would be to punch as a block of entries on a single card (or group of cards acting as a unit) first, the encoded representation of a compound's molecular structure followed by data as to its melting point, boiling point, refractive index, density, etc. If one of the properties were the solubility of some substance liquid at ordinary temperatures, - for example, ether - then additional conventions may be necessary to avoid ambiguity as to whether the solubility recorded by punching on our card refers to our liquid substance dissolved in water or water dissolved in our liquid substance. Such conventional methods of punching are relatively easy to set up as long as the possibility of ambiguity is restricted to solubility.

Solubility is only one of many relations in which ambiguity may be involved. Other examples might be the starting point and destination of a trip, temporal sequence of two or more events, differences in properties between substances such as "B is harder than A". This list of possibilities is capable of almost indefinite expansion.

Such relations in which there is no a priori certainty that A stands in relation R to B rather than B standing in relation R to A we shall term asymmetric relations.

The relation involved in many interactions may be asymmetric in nature. We have already mentioned the example of A dissolves B and its counterpart B dissolves A. An example from international relations might be A attacks B and B attacks A. In the realm of business we might have Company A is a subsidiary of Company B and as the possible alternative Company B is the subsidiary of Company A.

In considering means for resolving ambigous relations, it is helpful to use a simple notation. Thus, in the examples given above, we have used the letters A and B to refer to pairs of substances, countries and companies. If now we use R to indicate any one of the relations referred to, we might reduce our three pairs of asymmetric relations involving (i) solubility, (ii) attacking, and (iii) subsidiary status to a single pair of generalized relations:

A R's B

and

B R's A

In contemplating the problem of asymmetric relations from the viewpoint of machine searching methods, one very important consideration relates to various means for so expressing the relationship actually existing in a given instance so that the machine can take cognizance of that existing relationship when conducting searching and selecting operations.

Perhaps the simplest - though not necessarily the most readily usable - device for resolving asymmetric relations is the one employed very extensively by the English language. This method is to establish the convention that when A is in the relation R to B then this fact is indicated by the order of citation of the three symbols. Thus, for our example, we might stipulate that A shall be cited first followed by R and B in that order.

It might be observed that, as a matter of logic, or so far as machine searching is concerned there is no reason, except perhaps similarity to the familiar practices of the English language, why any one sequence of the three symbols should be given preference in establishing the convention selected to indicate that A is in relation R to B. Instead of A R B, we might with equal logic select B R A or A B R or B A R or R A B or R B A. Nor is it necessary that the inverse order of symbols be used to designate an inverse relationship. Thus we might use A B R to indicate that A is in relation R to B while using R A B to indicate that B is in relation R to A.

What is still important is that selections must be made and adhered to without exception, if order of citation of the symbols is to be used effectively to resolve ambiguity when dealing with asymmetric relations. For any given relation R, which is asymmetric, it would be entirely possible for the code dictionary to specify which order of symbols has been selected to designate that one entity is in relation R to another.

Whether this method is used or not depends, of course, on practical considerations. One of these will be the amount of difficulty experienced by the indexer and the encoder in establishing the proper sequence of symbols to indicate the relationship which is specified in the document being analyzed. Another consideration is the ability of the searching machine to discriminate one sequence of symbols from another. As already noted, machines now being developed have the needed discriminating ability and the decision as to whether order of citation of symbols will be used to resolve asymmetric ambiguity must be based on other considerations.

If we were never concerned with relations more complex than those involving three elements, as exemplified by A R's B, then adopting an ordering convention would perhaps afford the simplest and most advantageous solution of the problem. However, there are asymmetric relations of higher orders of complexity. For example, we may have relations involving four elements such as (1) A gives B to C or (u) B and C interact to produce A (as in plant hybridizing or in chemical relations). Such relations may be linked together as in barter trading with A giving B to C and C giving D to A. Or, to give another example, B and C may interact to produce A and D. To cope with such relations by an ordering convention involves either (1) much more intricate rules than required for the simple three-element relation A R's B or (2) the more complex relations must be broken up into smaller units. Thus we might establish the order A R B C to indicate A gives B to C (with R indicating the giving relation). Alternately, we might express this relation as the sum of A gives B and C receives B. This alternative involves a repetition of the symbol C and also double encoding of the relation (gives, receives) involved. In a more complex situation, the degree of repetition becomes so great as to raise grave questions as to the practicality of this approach .

In considering the advisability of relying on order as the sole means for indicating relationships between such elements as A, B, and R, it is a matter of considerable practical importance that in general conventions based on ordering will require the symbols to be arranged in a different sequence than that in which the corresponding elements appear in the document being analyzed. Thus, to cite a simple example, if the idea to be encoded is, "man bites dog", then this simple relationship may be expressed in the document by the sentence, "a dog

was bitten by a man". Rearranging this simple sentence is, of course, relatively easy. But as the coding conventions relating to ordered sequence become more complex - as is inevitable when handling more complex interactions - then following these conventions when encoding will almost certainly be difficult and time-consuming.

As is evident, many difficulties and complexities are encountered when the attempt is made to use order of citation of index entries as the sole means for resolving ambiguities caused by asymmetric relations. For this reason, it appears worth while to investigate other possibilities of taking asymmetric relations into account when conducting searches.

By taking a slightly different approach it is possible to use a predetermined order of citation as the basic means for resolving ambiguity involving asymmetry and also to avoid some of the difficulties mentioned in the preceding paragraphs. This approach might be described - in order to make it more readily understandable - as consisting of two steps. The first step requires that the index entries - perhaps most conveniently in encoded form - be arranged in the order which has been established as standard for resolving the inherent ambiguity of the asymmetric relationship. The second step is to attach numbers in the usual arithmetical order to the successive index entries (or to their encoded designations) in the ordered array. Actually in performing the encoding operation the first step would not be absolutely necessary. It would be necessary to keep in mind, however, which place in the array would be occupied by a given entry - or its encoded designation - even though the array itself were perhaps not set up in explicit form. In other words, it might suffice-particularly in simple cases - to have in mind which position a given code designation would occupy in an array and then assign it the appropriate number without going to the trouble of actually writing out the array itself.

This approach is closely akin to the concept of "argument positions" of formal logic. When a given relationship word, for example a transitive verb, requires citation of two other entities to make a complete statement concerning the action, then the relationship word is said to have two arguments. Our example of A R's B involves a relation R having two arguments A and B. Our example of A giving B to C involves a relation - namely giving - having three arguments.

If the approach under consideration were adopted, the code dictionary would be built up in such a way that entering a relationship term in the dictionary would involve not only its code designation, but also specification of the numerical indexes to be attached to the respective arguments.

It is instructive to consider how this approach may be applied to examples already discussed. Thus if the sequencing order A R B is taken as the basis for expressing that A stands in relation R to B, then the roles of A R and B in that relationship might be symbolized by attaching the prefix 1 to A, the prefix 2 to R and the prefix 3 to B. These new composite symbols may now be cited in any order without any fear of ambiguity as a result. Thus we might indicate that A stands in relation R to B by symbolic arrays of which 1A 2R 3B is one example. Another is 2R 1A 3B; another is 3B 1A 2R, etc.

If this approach is applied to our example of A and B reacting to form C and D we might base the numbering of the elements on the sequence A B R C D in which case - after applying the index numerals- we would have 1A 2B 3R 4C 5D. Similarly if E gives F and G to H we might take the sequence E R F G H as the basis and apply the index numerals to arrive at the symbolic representation, 1E 2R 3F 4G 5H. Certain disadvantages inherent in this approach become evident on considering these examples. First, the symbolic representation 1A 2B 3R 4C 5D attaches different numbers to A and B, even though those symbols both represent initial reacting substances, while the substances formed are also represented by symbols to which different numerical indexes are attached. In the other example, 1E 2R 3F 4G 5H, two different numeral indexes are attached to F and G even though both were given to H.

The first step toward improving this approach is to specify that the same role indicator shall be used with all entities that have same role. From the viewpoint of theoretical logic, this would mean that numerical indexes are used to indicate the argument type rather than the position of the symbol in a standardized array set up for the purpose of providing a basis for resolving ambiguity due to some asymmetric relation. If this is done, the symbolic representation of A and B reacting to form C and D might become:

1A 1B 2R 3C 3D

Similarly the representation of E giving F and G to H might take into account the identify of the roles of F and G by the notation:

1E 2R 2F 2G 3H

If the symbolism for indicating roles were set up in this way, compilation of the code dictionary would require each dictionary entry denoting a relation to specify the appropriate numerical indexes to be used to indicate the various roles associated with a given relation (be it "reacting", "giving", etc.).

These examples also point the way to a possibility for simplifying the use of numerical indexes - or similar symbols - to indicate

the diffeernt roles associated with various relations.

As already noted, earlier in this chapter, our analysis of asymmetric relations can be related to the "argument positions" concept of formal logic. This concept permits us to use a general symbolic designation to embrace a wide range of relationships. Thus R was used to designate four different relations while A and B denoted four different types of entities, as follows:

A	R's	B
Substance A	dissolves	substance B
Country A	attacks	country B
Company A	is subsidiary to	company B
Material A	is harder than	material B

The index 1, 2 and 3 might then be attached to A, R and B, respectively, when confronted with any one of the four relations denoted by the general symbolism A R's B.

Generalizing from this example, we can - if we deem it appropriate - group together and express by generalized symbolism relations which have the common feature of having the same number of logical arguments. The same role indicators can then be used with any set of relations.

It is perhaps obvious that we are under no compulsion to group together all relations which have the same number of logical arguments. One of the problems of code construction is to arrive at decisions as to how groupings of relations can be set up to best advantage so as to keep the system of role indicators as simple as possible and yet provide the type of discrimination effective in selecting needed information.

If role indicators are set up on the basis of groupings of relations characterized by the same number of logical arguments, it would scarcely be possible to ascribe any specific meaning to the role indicators which in fact do no more than provide means for resolving ambiguity arising from asymmetric relations. The possibility exists, however, to ascribe definable meaning to the role indicators. Thus in chemistry we might use the symbol "s" to denote a starting material and "p" to denote a reaction product. Thus we might symbolize the reaction of A and B to produce C and D by:

sA sB R pC pD

where R denotes chemical reaction. If it seemed appropriate we might generalize the symbol "s" to denote a wider range of entities, including, for example, plants used for hybridizing, while "p" might similarly be generalized to include entities produced, including, for

example, the result of hybridization. We might denote the hybridization of plants A and B to produce the hybrid C by the symbolism:

$$sA \; sB \quad H \quad pC$$

where H denotes hybridization. If the new variety C of a given species were obtained by other means--e.g., as a spontaneous mutation, or by plant selection--then the newly obtained plant would, by this approach, also be denoted by pC. A machine search directed to pC would then locate all those documents in which the plant variety C had been produced regardless of the means employed to establish the new variety. Such a search would exclude documents in which variety C had been used in some other role, e.g., parent plant used for hybridizing.

So far discussion in this appendix has centered on asymmetric relations and means for resolving ambiguities associated with them. In terms of formal logic we have been concerned with relations and their arguments, especially interactions and the entities directly concerned. Circumstances surrounding an interaction have been left out of consideration. Thus, for example, in speaking of chemical reactions no mention was made of temperature, pressure, inert solvents, catalysts and the like.